DOCUMENT RESUME

ED 190 597 TM 800 385

AUTHOR Huynh, Huynh; Saunders, Joseph C., III
TITLE Bayesian and Empirical Bayes Approaches to Setting Passing Scores on Mastery Tests. Publication Series in Mastery Testing.
INSTITUTION South Carolina Univ., Columbia. School of Education.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
REPORT NO RM-79-2
PUB DATE Apr 79
GRANT NIE-G-78-0087
NOTE 17p.; Paper presented at the Joint Annual Meetings of the American Educational Research Association and the National Council on Measurement in Education (San Francisco, CA, April 8-12, 1979).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Bayesian Statistics; *Cutting Scores; Grade 3; *Mastery Tests; Minimum Competency Testing; Primary Education; *Scoring Formulas; True Scores
IDENTIFIERS Binomial Error Model; Comprehensive Tests of Basic Skills; South Carolina Statewide Testing Program; Test Length

ABSTRACT

The Bayesian approach to setting passing scores, as proposed by Swaminathan, Hambleton, and Algina, is compared with the empirical Bayes approach to the same problem that is derived from Huynh's decision-theoretic framework. Comparisons are based on simulated data which follow an approximate beta-binomial distribution and on real test results from the Comprehensive Tests of Basic Skills administered in the South Carolina Statewide Testing Program. Both procedures lead to setting identical or almost identical passing scores as long as the test score distribution is reasonably symmetric or when the minimum mastery level or criterion level is high. Larger discrepancies tend to occur when this level is low, especially when the distribution of test scores is concentrated at a few extreme scores or when the frequencies are irregular. However, in terms of mastery/nonmastery decision, the two procedures result in the same classifications in practically all situations. The empirical Bayes procedures may be used for tests of any length, while the Bayesian procedure is recommended only for tests of eight or more items. Further, the empirical Bayes can be generalized and applied to more complex testing situations with less difficulty than the Bayesian procedure. (Author/CP)

PUBLICATION SERIES IN MASTERY TESTING
University of South Carolina
College of Education
Columbia, South Carolina 29208

Research Memorandum 79-2
April, 1979

BAYESIAN AND EMPIRICAL BAYES APPROACHES TO SETTING PASSING SCORES ON MASTERY TESTS

Huynh Huynh
Joseph C. Saunders III

University of South Carolina

Presented at the symposium "Psychometric approaches to domain-referenced testing" sponsored jointly by the American Educational Research Association and the National Council on Measurement in Education at their annual meetings in San Francisco, April 8-12, 1979.

ABSTRACT

The Bayesian approach to setting passing scores as proposed by Swaminathan, Hambleton, and Algina is compared with the empirical Bayes approach to the same problem that is derived from Huynh's decision-theoretic framework. Comparisons are based on simulated data which follow an approximate beta-binomial distribution and on real test data sampled from a statewide testing program. It is found that the two procedures lead to setting identical or almost identical passing scores as long as the test score distribution is reasonably symmetric or when the minimum mastery level or criterion level is high. Larger discrepancies tend to occur when this level is low, especially when the distribution of test scores is concentrated at a few extreme scores or when the frequencies are irregular. However, in terms of mastery/nonmastery decisions, the two procedures result in the same classifications in practically all situations. However, the empirical Bayes procedure may be used for tests of any length, while the Bayesian procedure is recommended only for tests of 8 or more items. Additionally, the empirical

This work was performed pursuant to Grant No. NIE-G-78-0087 with the National Institute of Education, Department of Health, Education, and Welfare, Huynh Huynh, Principal Investigator.

Bayes procedure can be generalized and applied to more complex testing situations with less difficulty than the Bayesian procedure.

1. INTRODUCTION

Among the many decision-theoretic approaches to setting passing scores (or standards) for mastery tests, there are at least two methods which rely on test data collected from a group of examinees. The Bayesian procedure, as presented in Swaminathan, Hambleton, and Algina (1975), assumes that prior knowledge regarding the examinees is exchangeable (Novick, Lewis, & Jackson, 1973) and can be quantified in some appropriate manner. On the other hand, the empirical Bayes approach, as formulated in Huynh (1976a), uses only the true ability distribution of the examinees and makes no assumption regarding prior knowledge about the examinees. Both procedures use test data collected from a group of examinees and establish passing scores for mastery tests by minimizing certain loss functions. The purpose of this paper is to present a comparison of the two sets of standards (passing scores) formulated under a variety of conditions which can be expected to be encountered in mastery testing or in minimum competency testing. The comparison will be made first on the basis of approximate beta-binomial test scores. Further comparisons will be made using the Comprehensive Tests of Basic Skills (CTBS, 1973) data collected in the 1978 South Carolina Statewide Testing Program.

2. AN OVERVIEW OF THE BAYESIAN AND EMPIRICAL BAYES APPROACHES

Overall Framework

The Bayesian framework as presented by Swaminathan et al. and the special empirical Bayes procedure described in Huynh (1976a, p. 70-73) start with a typical four-corner setup used in decision theory. (See Figure I, p. 16, for the basic elements of this setup.) Let θ (π in the notation of Swaminathan et al.) be the true score (or

true ability) of an examinee and x be the observed test score as obtained from an n-item test. For the binomial error model adopted in both standard setting approaches, θ is the proportion of items in a real or hypothetical item pool that an examinee answers correctly. Let a person be called a master if that person's true score θ is such that $\theta \geq \theta_o$ and a nonmaster if $\theta < \theta_o$. Here, θ_o is a given constant which defines the lower boundary of the mastery level or the criterion level. Since a person's true score cannot be observed directly, decisions about whether to call the person a master must be based on an observed test score. What remains to be determined is the cutoff score c that will be in some sense optimal.

On the basis of the test score x, a person is called a master if $x \geq c$ and a nonmaster if $x < c$. A correct decision is made whenever either (a) $\theta \geq \theta_o$ and $x \geq c$, or (b) $\theta < \theta_o$ and $x < c$. Otherwise, either a false positive error ($\theta < \theta_o$ and $x \geq c$) or a false negative error ($\theta \geq \theta_o$ and $x < c$) is encountered.

In the case where the loss associated with each error is constant, generality is not diminished if we let the loss incurred by a false positive error be equal to 1 and that associated with a false negative error be equal to Q. Here, Q expresses the ratio of the false negative error loss to the false positive error loss. (In the notation of Swaminathan et al., $Q = \ell_{21}/\ell_{12}$.)

Bayesian Approach

Now let an n-item test be given to m examinees. In the Bayesian procedure as implemented by Swaminathan et al., the prior information regarding the examinees is assumed to be exchangeable (i.e., prior knowledge regarding one examinee can be interchanged with that associated with another examinee without causing any disturbance in the decision problem). The model requires knowledge (prior belief) of the distribution of the variance of true scores for the group. (In point of fact, an arcsine transformation of θ is used.) This prior distribution is taken to be the inverse chi-square distribution with parameter λ and degrees of freedom ν. A recommended choice of ν is 8 (Novick, et al., 1973).

To assess λ, let t be the number of test items which would need to be administered to a typical examinee in order to obtain as much information about that examinee's θ as we already have. Then, $\lambda = 3/(2t+1)$. Wang (1973) has tables to facilitate computation in this procedure. In the setup of the Wang tables, λ/ν is chosen as .01, .02, .03, .04, and .05. These ratios correspond to the t values of 18.25, 8.875, 5.75, 4.1875, and 3.25. Given the prior information as revealed through λ and ν and the test data of m subjects, it is possible via the Wang tables to compute the two expected losses: $Pr(\theta < \theta_o \mid \text{test data})$ and $Q \cdot Pr(\theta \geq \theta_o \mid \text{test data})$ at each test score. A Bayesian passing score is then the smallest score at which the first expected loss is smaller than the second one. More details may be found in Swaminathan et al. (1975) and in Novick et al. (1973).

Empirical Bayes Approach

The empirical Bayes solution assumes that the m examinees constitute a random sample from a population for which the true ability θ follows a known distributional form such as the beta density with parameters α and β (Keats & Lord, 1962, page 68). Sample test data are used to obtain the estimates $\hat{\alpha}$ and $\hat{\beta}$, and the results are used to compute the probability of a false positive decision $Pr(\theta < \theta_o, x \geq c)$ and of a false negative decision $Q \cdot Pr(\theta \geq \theta_o, x < c)$ at a given cutoff score c. The optimum passing score (henceforth referred to simply as the passing score) will be the value of c at which the average loss, $Pr(\theta < \theta_o, x \geq c) + Q \cdot Pr(\theta \geq \theta_o, x < c)$, is the smallest.

The procedure is implemented as follows. Let $\bar{x}$ and s be the mean and standard deviation of the test scores, and let the Kuder-Richardson reliability coefficient be defined as

$$\hat{\alpha}_{21} = \frac{n}{n-1}\left(1 - \frac{\bar{x}(n-\bar{x})}{ns^2}\right).$$

Then

$$\hat{\alpha} = (-1 + 1/\hat{\alpha}_{21})\bar{x}$$

and

$$\hat{\beta} = -\hat{\alpha} + n/\hat{\alpha}_{21} - n.$$

For test scores with insufficient variability, $\hat{\alpha}_{21}$ may be negative. If this occurs simply replace $\hat{\alpha}_{21}$ by the smallest positive reliability estimate which happens to be available. Let I denote the incomplete beta function as tabulated in Pearson (1934) and implemented via computer programs such as the IBM Scientific Subroutine Package (1971) of the IMSL (1977). Then the passing score is the smallest integer c, at which

$$I(\hat{\alpha}+c, n+\hat{\beta}-c; \theta_o) \leq Q/(1+Q). \quad (1)$$

A normal approximation is available if there is a sufficiently large number of items and if θ_o is not near 0 or 1. Let ξ denote the $100/(1+Q)$ percentile of the unit normal distribution. Then the test passing score is nearly equal to

$$c = (n+\hat{\alpha}+\hat{\beta}-1)\theta_o + \xi\left[(n+\hat{\alpha}+\hat{\beta}-1)\theta_o(1-\theta_o)\right]^{\frac{1}{2}} - \hat{\alpha} + .5. \quad (2)$$

The data presented in Huynh (1976b) indicate that the passing score computed from Equation (2) does not differ appreciably from the one deduced from Inequation (1) when the test consists of 20 items and when θ_o is within the range from .50 to .80.

3. A COMPARISON OF BAYESIAN AND EMPIRICAL BAYES PASSING SCORES FOR APPROXIMATE BETA-BINOMIAL TEST DATA

The passing score obtained via the empirical Bayes approach, as revealed by Inequation (1), is based on test score data that follow a beta-binomial distribution. It may be of interest to compare the Bayesian approach to setting a passing score with the empirical Bayes approach, using test data which follow closely a beta-binomial form.

Both the present comparison and the one detailed in the next section are based on tests with ten items. In these comparisons, the criterion or minimum mastery level is set at θ_o = .60, .70, and .80. The loss ratio is chosen to be Q = .25, .50, 1.00, and 2.00. (A loss ratio smaller than one indicates that a false positive error is less serious than a false negative error.) To compute a passing score via the Bayesian approach, it is necessary to specify

the ratio λ/ν or, equivalently, the quantity t as described in Section 2. It may be recalled that t may be interpreted as the number of "test items" which are believed to be as informative as the prior belief about the examinees. In practical situations involving standard setting, it seems unreasonable to let the prior belief ν carry as much weight as the objective test data. In other words, it is unlikely that t is too close to n. Thus for the comparisons based on 10-item tests reported in this section and in Section 4 as well as the comparisons based on 20-item tests described in Section 5, the t-values are chosen to be 8.875 (λ/ν = .02), 5.75 (λ/ν = .03), 4.1875 (λ/ν = .04), and 3.25 (λ/ν = .05).

The first five test score frequency distributions (labeled A1 through A5 in Table 1) serve as the data base for the comparison of the passing scores computed by the two procedures using test score distributions that are approximately beta-binomial. Each is deliberately chosen (i) to yield an s_g^2 value (variance of the arcsine-square-root transformation of the test scores) conforming as closely as possible to the tabulated s_g^2 values of the Wang tables (so that no interpolation would be necessary) and (ii) to reflect several degrees of skewness and variability thought to be typical of mastery testing situations. (Also in Table 1, and explained below, are distributions of actual test scores from the South Carolina Statewide Testing Program.) It may be noted that in Table 1, the quantity D(%) represents the maximum percent difference between the observed and beta-binomial-fitted cumulative frequencies. A small D-value indicates a good fit.

Table 2 reports the Bayesian passing scores and the corresponding empirical Bayes passing scores *(in italics)* for several combinations of θ_o, Q, and t. The data indicate that for the situations under consideration, the Bayesian and empirical Bayes passing scores are identical, or nearly so, as long as the test score distribution is reasonably symmetrical (Cases A2, A4, and A5). For highly skewed distributions (Cases A1 and A3) the two passing

TABLE 1

Frequency Distributions of Test Scores Used in Comparisons of Passing Scores

Data Set	Source/ Subtest	m*	D(%)†	S.D.	Skewness	Frequency at score of 0	1	2	3	4	5	6	7	8	9	10
	Approximate Beta-Binomial															
A1	Fictitious	40	3.1	1.36	-0.61						1	3	6	8	11	11
A2	Fictitious	80	1.0	1.87	-0.31			1	3	6	10	13	16	15	11	5
A3	Fictitious	40	1.2	1.01	-1.51							1	2	4	10	23
A4	Fictitious	40	1.6	2.01	-0.02		1	3	5	6	7	7	5	4	2	0
A5	Fictitious	40	1.0	2.15	0.12	1	3	5	6	7	6	5	4	2	1	0
	Comprehensive Tests of Basic Skills															
B1	Mathematics concepts and applications	20	6.7	1.28	-0.63							2	1	6	4	7
B2	Mathematics computations	20	9.2	1.45	-0.24							3	4	3	4	6
B3	Spelling	20	6.1	1.76	-1.04					2	0	1	2	6	4	5
B4	Social studies	40	6.2	2.11	0.27		1	4	5	9	5	5	6	3	1	1
B5	Language expression	40	8.2	1.86	-0.53			1	1	5	3	4	11	10	3	2
B6	Reading	40	4.1	1.22	-2.12						1	1	2	3	3	30
B7	Science	60	5.6	1.74	-0.22				2	6	10	8	14	8	12	0
B8	Reading vocabulary	60	3.2	1.56	-1.75				1	0	3	1	5	5	16	29
B9	Reading vocabulary	80	2.7	1.68	-1.49				2	1	2	5	6	11	23	30
B10	Spelling	80	2.1	1.50	-1.44				1	0	2	4	7	12	16	38

*m = total number of scores in the distribution.

†D(%) represents the maximum percent difference between the observed and beta-binomial-fitted cumulative frequencies. All are not significant at the ten percent level of significance.

scores rarely differ by more than one unit when the criterion level θ_o is relatively high (.70 or .80) and when λ/ν is such that t is not too close to n, say when λ/ν is at least .03. Large discrepancies, however, may occur at a low criterion level such as .60 or when t is close to n.

TABLE 2

Bayesian and Empirical Bayes Passing Scores for Five Approximate Beta-Binomial Test Score Distributions

Data Set	θ_o	Bayesian (at λ/ν = .02,.03,.04,.05) and empirical Bayes *(in italics)* at Q = .25	Q = .50	Q = 1.00	Q = 2.00
A1	.60	4, 5, 6, 6, *4*	3, 4, 5, 5, *2*	2, 3, 4, 4, *1*	1, 2, 3, 3, *0*
	.70	7, 8, 8, 8, *6*	6, 7, 7, 7, *5*	5, 5, 6, 6, *4*	4, 4, 5, 5, *3*
	.80	10,10,10,10, *9*	9, 9, 9, 9, *8*	8, 8, 8, 8, *7*	7, 7, 7, 7, *6*
A2	.60	7, 8, 8, 8, *7*	6, 7, 7, 7, *6*	5, 6, 6, 6, *5*	4, 4, 5, 5, *4*
	.70	10,10, 9, 9, *9*	9, 9, 9, 9, *9*	8, 8, 8, 8, *8*	7, 7, 7, 7, *7*
	.80	10,10,10,10,*10*	10,10,10,10,*10*	10,10,10,10,*10*	9, 9, 9, 9, *9*
A3	.60	1, 3, 4, 4, *3*	1, 2, 3, 3, *2*	0, 1, 2, 2, *1*	0, 1, 1, 2, *0*
	.70	4, 5, 6, 6, *6*	3, 4, 5, 5, *5*	2, 3, 4, 4, *4*	1, 2, 3, 3, *3*
	.80	8, 8, 9, 9, *8*	7, 7, 8, 8, *7*	5, 6, 7, 7, *6*	4, 5, 6, 6, *5*
A4	.60	9, 9, 9, 9, *9*	9, 8, 8, 8, *8*	8, 7, 7, 7, *8*	7, 6, 6, 6, *6*
	.70	10,10,10,10,*10*	10,10,10,10,*10*	10, 9, 9, 9,*10*	9, 9, 8, 8, *9*
	.80	10,10,10,10,*10*	10,10,10,10,*10*	10,10,10,10,*10*	10,10,10,10,*10*
A5	.60	10,10, 9, 9,*10*	9, 9, 9, 9, *9*	8, 8, 8, 8, *8*	7, 7, 7, 7, *7*
	.70	10,10,10,10,*10*	10,10,10,10,*10*	10,10, 9, 9,*10*	9, 9, 9, 9, *9*
	.80	10,10,10,10,*10*	10,10,10,10,*10*	10,10,10,10,*10*	10,10,10,10,*10*

4. A COMPARISON OF BAYESIAN AND EMPIRICAL BAYES PASSING SCORES FOR CTBS TEST DATA

This phase of the study is based on a 10% systematic sample of the entire third grade CTBS-Level C data file compiled during the 1978 South Carolina Statewide Testing Program. To obtain the frequency distributions labeled as B1 to B10 (in Tables 1 and 3), the following procedure was used. First, ten 10-item subtests were assembled by random selection of items from each CTBS subtest. Next, for each 10-item subtest, a frequency distribution was constructed for each school district which had at least 20 students in the systematic sample, and the corresponding s_g^2 value was obtained. (The s_g^2 values were distributed as follows: .10 to .50 (32%), .51 to .75 (38%), .76 to 1.00 (20%), and more than 1.00 (10%). Large s_g^2 values tended to associate with subtests dealing with reading comprehension (sentences or paragraphs), language expression, and language mechanics.) Third, among the frequency distributions with s_g^2 values included between .01 and .05, ten were finally selected

and altered slightly so that the total number of examinees (m) was exactly 20, 40, 60, or 80.

Table 3 lists the Bayesian and empirical Bayes passing scores under a variety of conditions. As in the previous section, the data

TABLE 3

Bayesian and Empirical Bayes Passing Scores for Ten CTBS Test Score Distributions

Data Set	θ_o	Bayesian (at λ/ν = .02,.03,.04,.05) and empirical Bayes *(in italics)* at Q = .25	Q = .50	Q = 1.00	Q = 2.00
B1	.60	5, 5, 6, 6, *3*	4, 4, 5, 5, *2*	3, 3, 4, 4, *1*	2, 2, 3, 3, *0*
	.70	7, 7, 8, 8, *6*	6, 6, 7, 7, *5*	5, 5, 6, 6, *4*	4, 4, 5, 5, *3*
	.80	10,10,10,10, *9*	9, 9, 9, 9, *8*	8, 8, 8, 8, *7*	7, 7, 7, 7, *6*
B2	.60	6, 6, 6, 6, *5*	5, 5, 5, 5, *4*	4, 4, 4, 5, *2*	3, 3, 3, 4, *1*
	.70	8, 8, 8, 8, *7*	7, 7, 7, 7, *6*	6, 6, 6, 6, *5*	5, 5, 5, 6, *4*
	.80	10,10,10,10, *9*	9, 9, 9, 9, *9*	8, 8, 8, 8, *8*	7, 7, 8, 8, *7*
B3	.60	6, 6, 7, 7, *6*	5, 5, 6, 6, *6*	4, 4, 5, 5, *5*	3, 4, 4, 4, *4*
	.70	8, 8, 8, 8, *8*	7, 7, 8, 8, *7*	6, 7, 7, 7, *6*	5, 6, 6, 6, *6*
	.80	10,10,10,10,*10*	9, 9, 9, 9, *9*	9, 9, 9, 9, *8*	8, 8, 8, 8, *7*
B4	.60	9, 9, 9, 9, *9*	9, 8, 8, 8, *8*	8, 8, 7, 7, *7*	7, 7, 6, 6, *7*
	.70	10,10,10,10,*10*	10,10,10,10,*10*	10, 9, 9, 9, *9*	9, 9, 8, 8, *9*
	.80	10,10,10,10,*10*	10,10,10,10,*10*	10,10,10,10,*10*	10,10,10,10,*10*
B5	.60	8, 8, 8, 8, *7*	7, 7, 7, 7, *6*	6, 6, 6, 6, *5*	4, 5, 5, 5, *4*
	.70	10,10, 9, 9,*10*	9, 9, 9, 9, *9*	8, 8, 8, 8, *8*	7, 7, 7, 7, *7*
	.80	10,10,10,10,*10*	10,10,10,10,*10*	10,10,10,10,*10*	9, 9, 9, 9, *9*
B6	.60	2, 3, 4, 5, *6*	1, 2, 3, 4, *6*	1, 2, 2, 3, *5*	0, 1, 1, 2, *4*
	.70	5, 5, 6, 7, *8*	3, 4, 5, 6, *7*	2, 3, 4, 5, *6*	2, 2, 3, 4, *6*
	.80	8, 8, 9, 9, *9*	7, 7, 8, 8, *8*	6, 6, 7, 7, *8*	4, 5, 6, 6, *7*
B7	.60	8, 8, 8, 8, *7*	7, 7, 7, 7, *6*	5, 6, 6, 6, *5*	4, 5, 5, 5, *4*
	.70	10,10,10,10, *9*	9, 9, 9, 9, *9*	8, 8, 8, 8, *8*	7, 7, 7, 7, *7*
	.80	10,10,10,10,*10*	10,10,10,10,*10*	10,10,10,10,*10*	10,10, 9, 9,*10*
B8	.60	3, 4, 5, 6, *6*	2, 3, 4, 5, *6*	2, 2, 3, 4, *5*	1, 2, 2, 3, *4*
	.70	6, 7, 7, 8, *8*	5, 6, 6, 7, *7*	4, 5, 5, 6, *6*	3, 4, 4, 5, *6*
	.80	9, 9, 9, 9, *9*	8, 8, 9, 9, *8*	7, 7, 8, 8, *8*	6, 6, 7, 7, *7*
B9	.60	4, 5, 5, 6, *6*	3, 4, 4, 5, *6*	2, 3, 3, 4, *5*	1, 2, 3, 3, *4*
	.70	7, 7, 8, 8, *8*	4, 6, 7, 7, *7*	4, 5, 6, 6, *6*	3, 4, 5, 5, *6*
	.80	9,10,10,10, *9*	9, 9, 9, 9, *9*	8, 8, 8, 8, *8*	6, 7, 7, 7, *7*
B10	.60	3, 4, 5, 6, *6*	2, 3, 4, 5, *5*	1, 2, 3, 4, *5*	1, 1, 2, 3, *4*
	.70	6, 7, 7, 8, *8*	5, 6, 6, 7, *7*	4, 4, 5, 6, *6*	3, 3, 4, 5, *5*
	.80	9, 9, 9, 9, *9*	8, 8, 9, 9, *8*	7, 7, 8, 8, *8*	6, 6, 7, 7, *7*

show that the two sets of passing scores are the same, or nearly so, as long as the test score distribution is reasonably symmetric (see cases B4, B5, and B7). Discrepancies in these situations are rarely larger than one unit. For most other situations, the difference between the two values for a passing score is seldom larger than one unit when the criterion θ_o is .70 or .80 and when λ/ν is at least .03. The same magnitude of difference, one unit, also tends to hold at θ_o = .60 unless the test scores pile up at extreme values (Case B6) or unless the frequencies are fairly irregular (Case B1).

5. ADDITIONAL DATA FOR MODERATELY SKEWED DISTRIBUTIONS

Additional comparisons were made for ten 20-item tests with distributions having skewness ranging from -1.109 to .117 (see Table 4). These tests were assembled in the same way as the 10-item tests described in Section 4. As in the previous sections, the criterion level θ_o was set at .60, .70, and .80, and the loss ratio Q at .25, .50, 1.00, and 2.00. The prior knowledge about the examinees was assumed to be equivalent to a number of items, t, of 8.875 (λ/ν = .02), 5.75 (λ/ν = .03), 4.1875 (λ/ν = .04), and 3.25 (λ/ν = .05). For all the 480 combinations under consideration, the

TABLE 4

Frequency Distribution of Scores on Ten CTBS Subtests Mentioned in Section 5

	Frequency at score of															
Subtest	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Reading vocabulary							1	1	5	3	4	7	4	8	3	4
Spelling								1	1	2	3	2	3	8	12	8
Science		1	1	1	3	3	4	3	1	9	4	5	2	1	1	1
Social studies	2	0	2	0	3	1	2	2	6	9	1	4	4	1	3	0
Social studies		1	2	5	3	3	1	6	5	4	2	2	5	0	0	1
Reading vocabulary			2	0	0	2	1	4	4	3	3	4	8	3	4	2
Mathematics concepts and application		1	0	0	1	2	3	2	3	4	0	7	7	2	6	2
Reading vocabulary								1	2	3	2	5	5	6	9	7
Social studies	1	3	1	1	1	0	2	5	3	6	3	5	4	4	1	0
Science	1	1	4	2	2	2	4	2	4	2	3	4	3	5	0	1

absolute value of the discrepancies between the two computed passing scores are distributed as follows: 0 (35%), 1 (37%), 2 (15%), 3 (5%), and 4 or more (8%). Hence in about three-fourths of all situations, the Bayesian and empirical Bayes passing scores do not differ from each other by more than one unit.

6. AGREEMENT OF MASTERY/NONMASTERY DECISIONS

As noted in Section 4, there are situations (such as some cases associated with the A1, B1, and B6 data sets) where the passing scores obtained from the two methods differ appreciably. This may seem disheartening. However, the procedures provide mastery/nonmastery classifications which are in high agreement for most cases under consideration. For Data Set A1 with θ_o = .60 and .70, for example, the combined proportions of students identically classified in either the mastery or nonmastery category by the Bayesian procedure (with λ/μ = .05) and by the empirical Bayes procedure are 88%, 95%, 99%, and 100% for Q = .25, .50, 1.00, and 2.00 respectively. Over the fifteen data sets of Table 1 and with the same values for λ/ν and Q, the proportions of identical classifications reach 94%, 96%, 98, and 97% respectively. As for the data of Table 4, these proportions stand at 98%, 98%, 98%, and 97%.

Though the overall agreement for classifications is high for the data considered in this study, some individual cases may show less agreement than others. These cases include situations such as A2 with θ_o = .60, Q = .25, and λ/ν = .05 where the Bayesian passing score of 8 and the empirical Bayes passing score of 7 are located near the center of the test score distribution. The shift of only one unit in test score in this case actually causes 16 students out of a total of 80 to be classified differently by the two procedures. Visible disagreement between the classifications defined by the Bayesian and empirical Bayes procedures may occur in situations where scores with high frequencies of occurrence are selected as the passing scores. If this is the case, the proportion of students classified in the mastery (or nonmastery) category is not likely to be close to either 0% or 100%. In other situations where

most students are declared masters (Data Set A1 with θ_o = .80, λ/ν = .05, and Q = 2.00) or nonmasters (Data Set A5 with θ_o = .70, λ/ν = .05, and Q = 1.00), the agreement in classifications is almost perfect.

7. DISCUSSION AND CONCLUSION

The results described in previous sections may be summarized as follows: (i) Bayesian passing scores and those computed via the empirical Bayes procedure are identical or almost identical as long as the test score frequency distribution is reasonably symmetric or when the criterion level θ_o is sufficiently high (.70 or .80); (ii) large discrepancies in passing scores may occur at criterion levels of .60 (or below), especially when the test scores pile up at a few extreme values or when the frequency distribution is irregular; (iii) however, mastery/nonmastery decisions derived from the two procedures are most often identical. Overall, the combined proportion of students similarly classified by both procedures is about 97%.

All in all, there is little difference between the Bayesian approach as described by Swaminathan et al. and the Huynh empirical Bayes procedure described here, either in terms of the resulting passing scores or in terms of the mastery/nonmastery categorization.

It should be pointed out that the procedure by Swaminathan et al. relies on a normal arcsine-square-root transformation of the test data and is therefore considered adequate only when the test has at least 8 items. In addition, the scheme requires the evaluation of certain posterior probabilities. This may be done via the MARPRO computer program (mentioned in Wang, 1973) or via the Wang tables. To the chagrin of the writers, many frequency distributions such as those derived from the CTBS test data of the South Carolina Statewide Testing Program have s_g^2 values much larger than the upper bound of .05 allowed in the above-mentioned tables. In addition, the constraint of having at least 8 items seems to be quite severe in many practical situations involving objective-

referenced testing. Such tests frequently have 5 or fewer items per objective.

The empirical Bayes approach in its simplest form, as presented in Huynh (1976a), requires that the test scores follow a beta-binomial distribution. There are indications (Keats & Lord, 1962; Duncan, 1974; Huynh & Saunders, 1979; also see Table 1) that the model adequately fits many test score distributions. Moreover, it is known (Subkoviak, 1978; Huynh & Saunders, 1979) that the model is useful in the estimation of the reliability of mastery classification based on one test administration. In addition, using the empirical Bayes approach, passing scores may be computed for tests of any length and can be approximated quickly via Equation (2).

It may be noted that the Bayesian and empirical Bayes procedures discussed in this paper deal with the setting of passing scores for a particular test. Both procedures assume the availability of a minimum mastery or criterion level θ_o and the availability of other information such as Q, the ratio of the loss incurred by a false positive decision to that incurred by a false negative one. In the context of testing for instructional purposes, θ_o may be based on the judgment of a curriculum specialist or a knowledgeable teacher and Q may be assessed via the time losses encountered by a misdecision (Huynh, 1976a). The issue is much more involved for end-of-program certification, such as high school graduation (minimum competency) testing programs legislated in several states. The reader is referred to Jaeger (1976) and Shepard (1976) for insight regarding some of these issues.

The empirical Bayes approach with the availability of a predetermined criterion level, however, is only the simplest form of the general framework of mastery evaluation as approached by Huynh (1976a). The essential component of this model is an external task (real or hypothetical) that examinees are supposed to perform once they are granted mastery of the objectives or content upon which a test is based. Such an external task may be identified in the context of instruction, especially when instructional units are

sequenced in some logical order. If this requirement is fulfilled, the specification of θ_o is no longer necessary. Some suggestions for solutions along this line have been presented elsewhere (Huynh, 1976a, p. 73-75; Huynh, 1977; Huynh & Perney, 1979). To the knowledge of the writers, the Bayesian approach as presented by Swaminathan et al. has not been generalized to situations other than those involving constant losses and when a criterion level is available. Although such a generalization may be made, the numerical analysis would be more involved than can be expected from the empirical Bayes approach.

As indicated previously, both standard setting procedures studied in this paper are based on group data and therefore are appropriate to the extent that minimization of loss is considered for the entire group of examinees. This may be the case for minimum competency testing where resources for remedial instruction are limited. Procedures relating to standard setting in the absence of group data are available (see, for example, Huynh, 1978).

In conclusion, the empirical Bayes approach yields mastery/nonmastery decisions identical in most cases to those based on the Bayesian approach. In addition, the former approach is simpler in terms of computations, is applicable to any test length, and has been generalized to more complex testing situations.

BIBLIOGRAPHY

Comprehensive Tests of Basic Skills, Level C (1973). Monterey, California: CTB/McGraw-Hill.

Duncan, G. T. (1974). An empirical Bayes approach to scoring multiple-choice tests in the misinformation model. Journal of the American Statistical Association 69, 50-57.

Huynh, H. (1976a). Statistical consideration of mastery scores. Psychometrika 41, 65-78.

Huynh, H. (1976b). On mastery scores and efficiency of criterion-referenced tests when losses are partially known. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 19-23.

Huynh, H. (1977). Two simple classes of mastery scores based on the beta-binomial model. Psychometrika 42, 601-608.

Huynh, H. (1978). A nonrandomized minimax solution for mastery scores in the binomial error model. Research Memorandum 78-2, Publication Series in Mastery Testing. University of South Carolina College of Education.

Huynh, H. & Perney, J. C. (1979). Determination of mastery scores when instructional units are linearly related. Educational and Psychological Measurement 39, 317-325.

Huynh, H. & Saunders, J. C. (1979). Accuracy of two procedures for estimating reliability of mastery tests. Research Memorandum 79-1, Publication Series in Mastery Testing. University of South Carolina College of Education. Also presented at the annual conference of the Eastern Education Research Association, Kiawah Island, South Carolina, February 22-24, 1979.

IBM Application Program, System/360 (1971). Scientific subroutines package (360-CM-03X) Version III, Programmer's manual. White Plains, New York: IBM Corporation Technical Publication Department.

IMSL Library 1 (1977). Houston: International Mathematical and Statistical Libraries.

Jaeger, R. M. (1976). Measurement consequences of selected standard-setting models. Florida Journal of Educational Research 18, 22-27.

Keats, J. A. & Lord, F. M. (1962). A theoretical distribution for mental test scores. Psychometrika 27, 59-72.

Novick, M. R., Lewis, C. & Jackson, P. H. (1973). The estimation of proportions in m groups. Psychometrika 38, 19-45.

Pearson, K. (1934). Tables of the Incomplete Beta Function. Cambridge: University Press.

Shepard, L. A. (1976). Setting standards and living with them. Florida Journal of Education Research 18, 23-32.

Subkoviak, M. J. (1978). Empirical investigation of procedures for estimating reliability of mastery tests. Journal of Educational Measurement 15, 111-116.

Swaminathan, H., Hambleton, R. K. & Algina, J. (1975). A Bayesian decision-theoretic procedure for use with criterion-referenced tests. Journal of Educational Measurement 12, 87-98.

Wang, M. M. (1973). Tables of constants for the posterior marginal estimates of proportions in m groups. ACT Technical Bulletin No. 14. Iowa City, Iowa: The American College Testing Program.

FIGURE I

Four Categories of Classification Based on Two Test Administrations

First Testing / Second Testing	Nonmastery	Mastery
Mastery	Nonmastery-Mastery	Mastery-Mastery (consistent decision)
Nonmastery	Nonmastery-Nonmastery (consistent decision)	Mastery-Nonmastery

ACKNOWLEDGEMENT

This work was performed pursuant to Grant NIE-G-78-0087 with the National Institute of Education, Department of Health, Education, and Welfare, Huynh Huynh, Principal Investigator. Points of view or opinions stated do not neccessarily reflect NIE positions or policy and no endorsement should be inferred. The editorial assistance of Anthony J. Nitko is gratefully acknowledged.